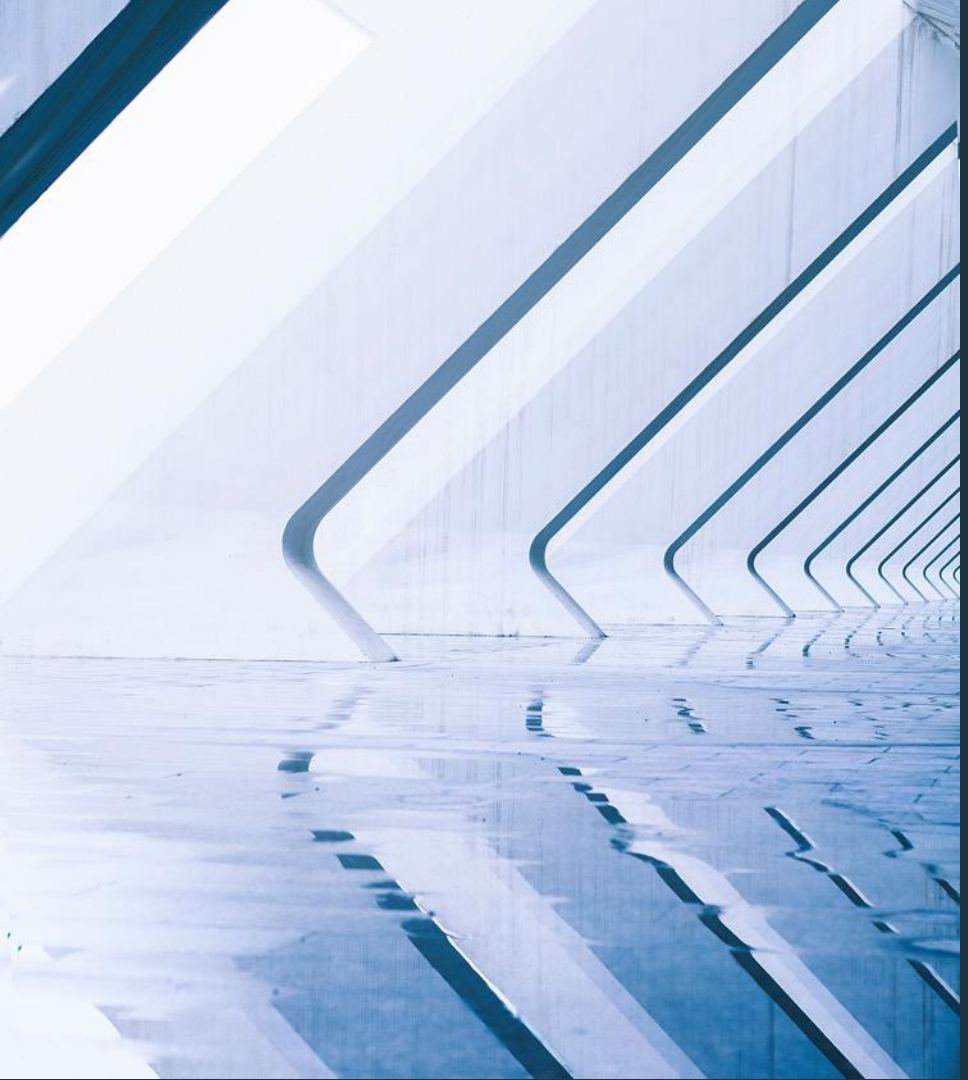
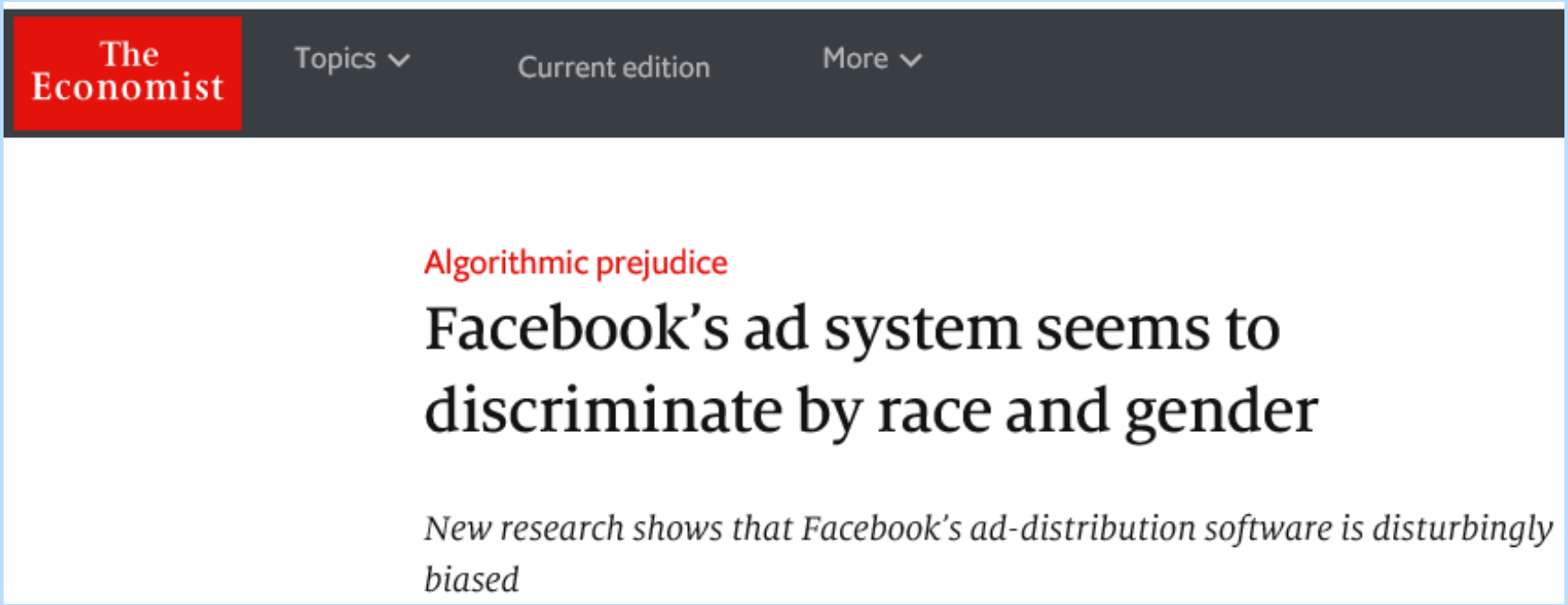


FICO

Bias In, Bias Out?

How to Fight Biases in AI and Create Trust

Dr. Gerald Fahner
Analytic Science-Sr Principal Scientist, FICO



THE DAILY NEWSLETTER
Sign up to our daily email newsletter

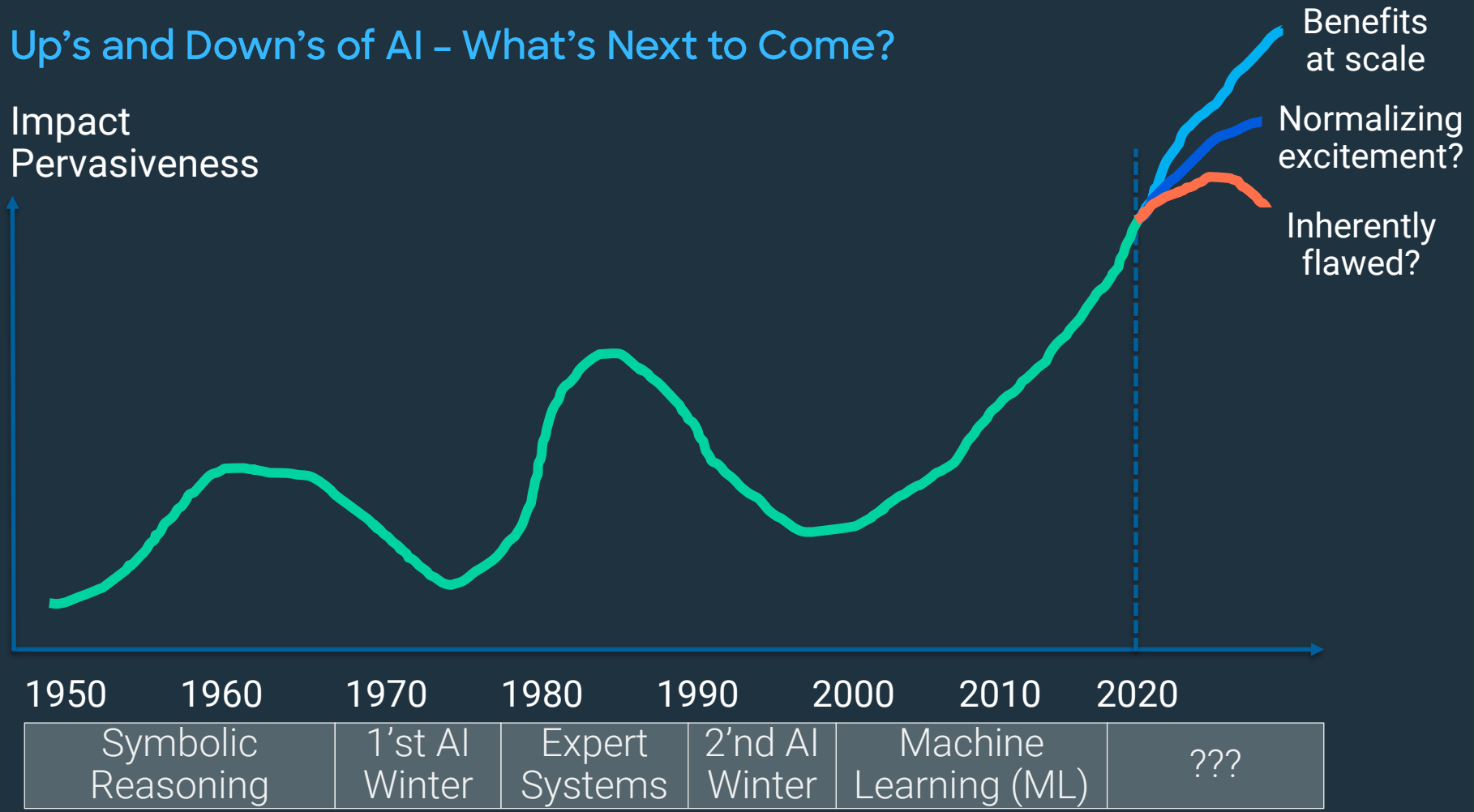
NewScientist

News Technology Space Physics Health Environment Mind Video | Travel Events Jobs

Face-recognition software is perfect – if you're a white man

TECHNOLOGY 13 February 2018

Up's and Down's of AI – What's Next to Come?

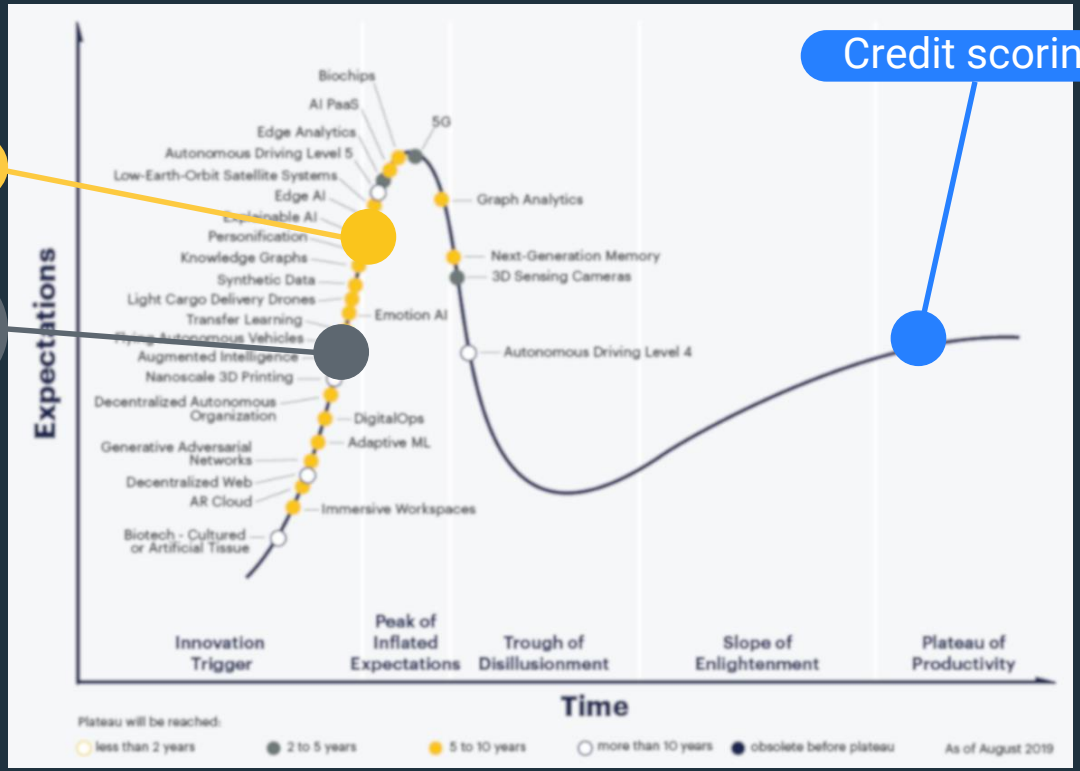


Gartner Hype Cycle for Emerging Technologies, 2019

Explainable AI

Augmented Intelligence

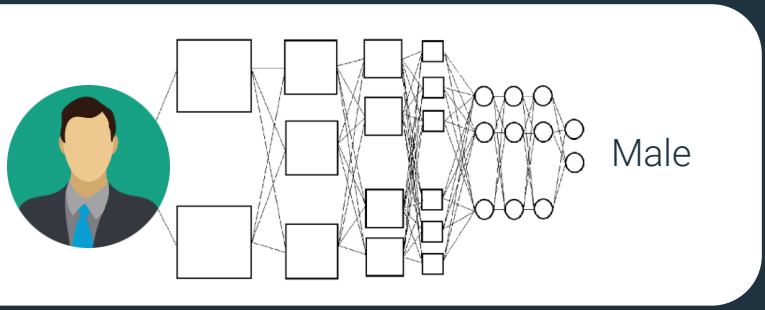
Credit scoring



Bias Culprit 1: Using Training Data That Lack Diversity

Training Data Labelled by Gender

Deep NN for gender classification



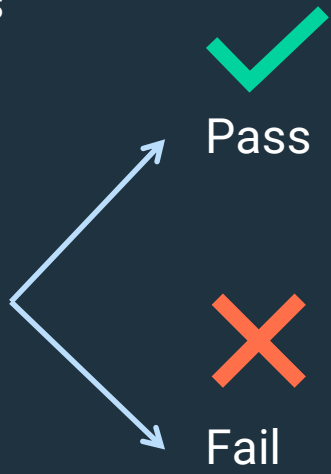
Error rates for various commercial products*	
Dark Females	24% to 36%
Light Females	0% to 2%
Dark Males	0% to 6%
Light Males	0.0% to 0.3%

*J. Buolamwini and T. Gebru, Gender Shades: "Intersectional Accuracy Disparities in Commercial Gender Classification", Proceedings of Machine Learning Research 81:1–15, 2018

Bias Culprit 2: Training a Machine With Subjective Labels

Resumes

Judgmental prescreen decisions



Bias Culprit 3: Letting Irrelevant “Big Data” Features Creep Into a Model

Resumes

Zillion features

	innovate	softball	influence	women	+ many more
Resume 1	1	0	0	1	…
Resume 2	0	0	1	2	…
Resume 3	3	1	0	0	…

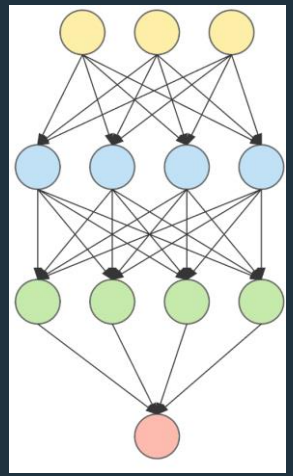
Bias Culprit 4: Prioritizing “Best Fit” Over Transparency

Biased

Historic data (features and labels)

Train model

Zillion features



Pass/Fail

Biased

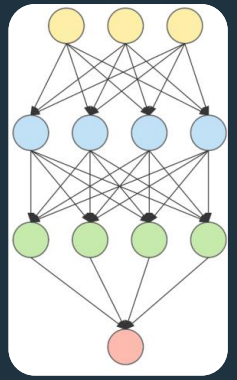
Feature	Impact on ‘Pass’ decision
innovate	+1.37
softball	-0.22
influence	+0.28
women	-0.31
execute	+0.57
…	…

Vicious Artificial Intelligence Cycle

ML “best fit”

Biased data, abundance of irrelevant features

Learning unpalatable associations



Biases are amplified

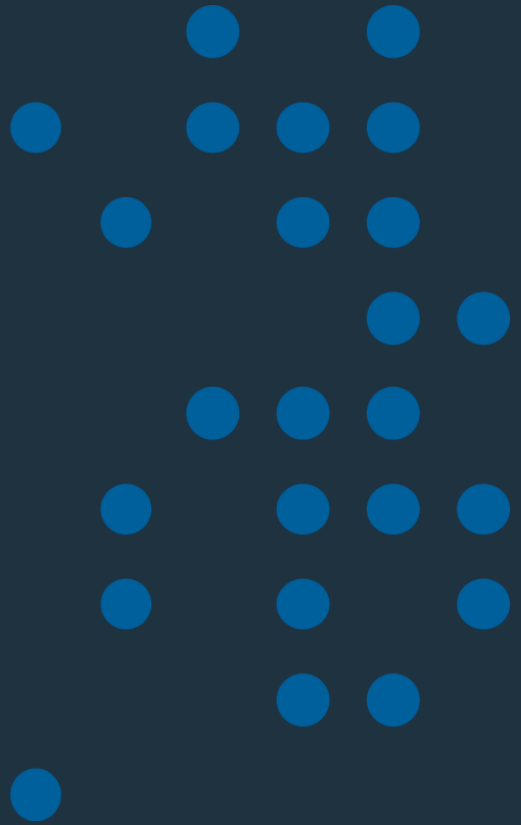
Biased, misguided decisions

Questions for You

How can we mitigate algorithmic bias (discrimination) against demographic groups?

A. Use all available data sources to build the most accurate model

B. Don’t use demographic group membership as a predictor



In the US, if a credit scoring model:

Doesn’t use prohibited factors or proxies for prohibited factors

AND

Is empirically derived, demonstrably and statistically sound (“EDDSS”)

THEN

The model affords certain protections against disparate treatment and disparate impact claims

Questions for You

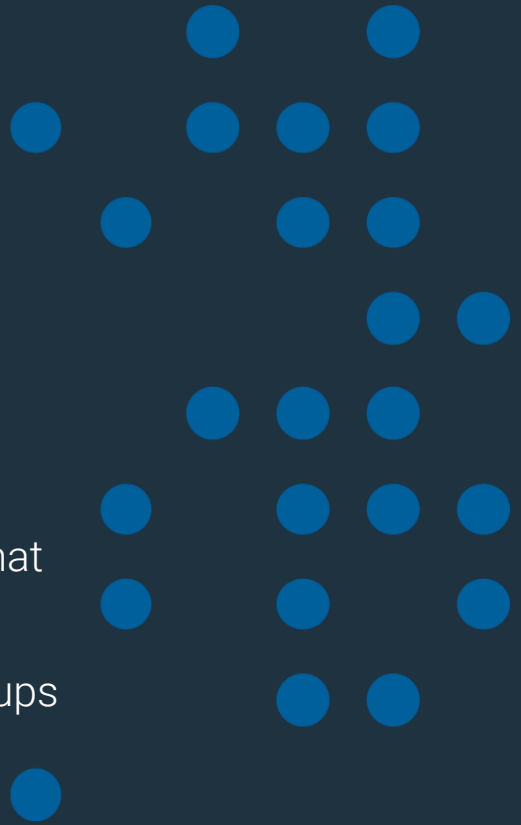
How can we mitigate algorithmic bias (discrimination) against demographic groups?

A. Use all available data sources to build the most accurate model

B. Don't use demographic group membership as a predictor

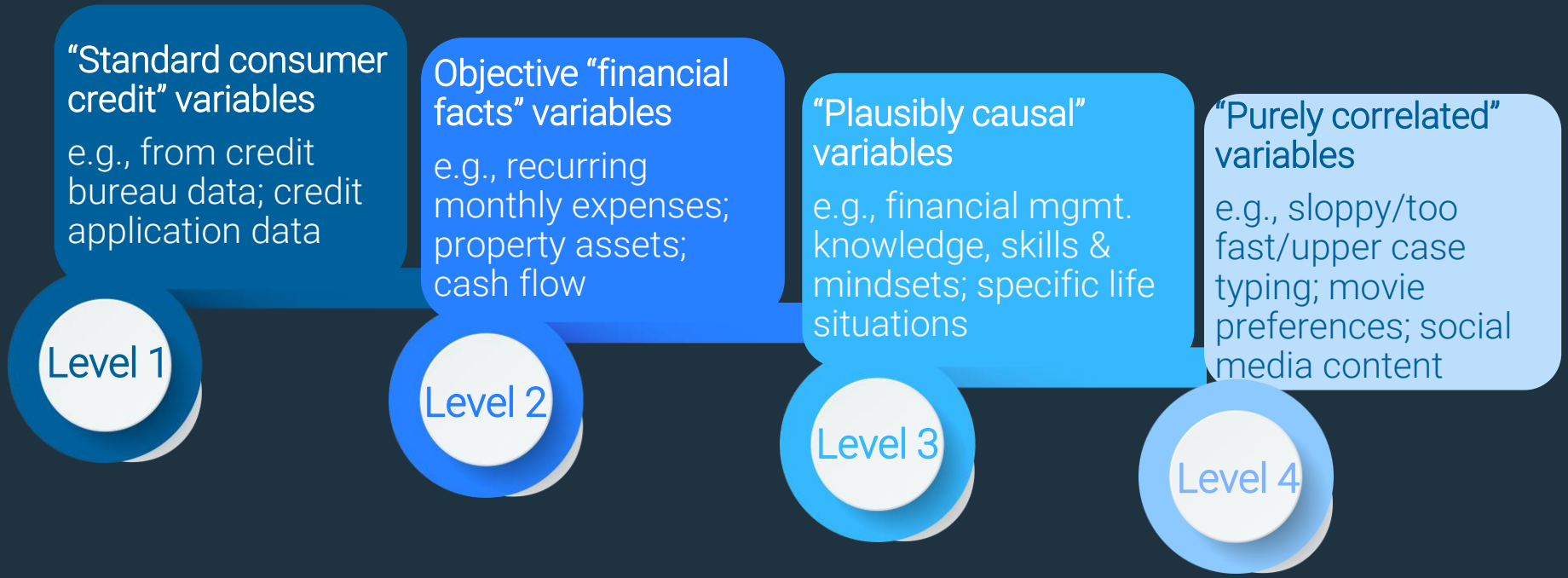
C. Allow score cutoffs to differ between demographic groups, such that equal proportions from each group are selected

D. Collect relevant and reliable variables that are predictive for all groups



Collect relevant and reliable variables that are predictive for all groups

Bias Buster 1: Prioritize Plausibly Causal Variables Predictive for All Groups



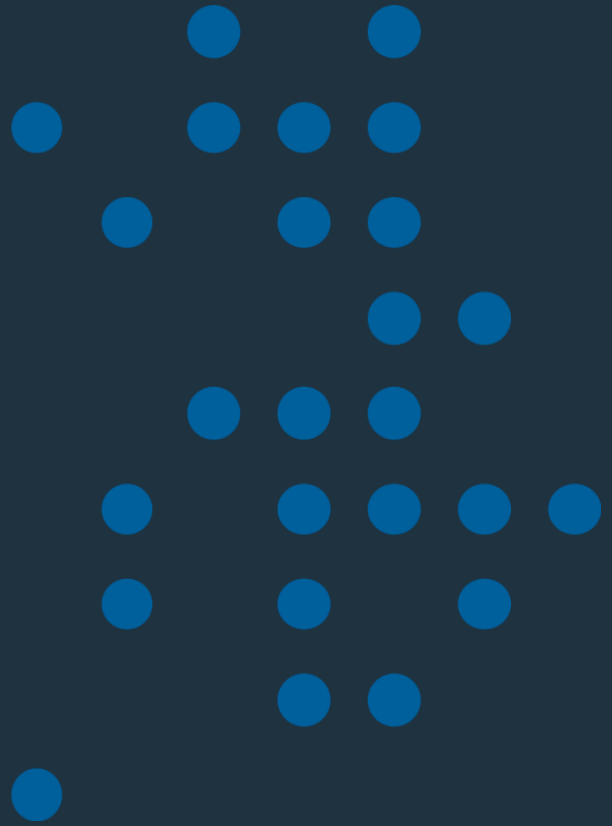
Questions for You

Do you believe the following situations are fair?

A. A parole algorithm is equally predictive for black and white defendants

B. A parole algorithm falsely flags black defendants twice as likely as future criminals, than white defendants

C. A university admits 44% of male applicants but only 35% of female applicants



Berkeley Admissions Decisions*

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

Biased against women?

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

Not biased on department level

Women applied more to departments that are hard to get into

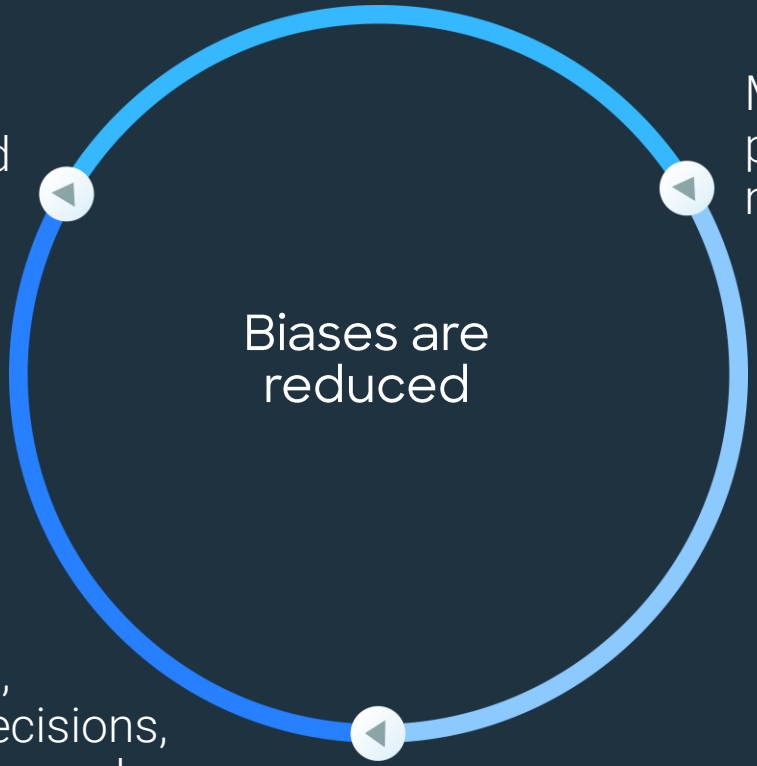
*P. Bickel, E. Hammel and J. O'Connell: "Sex Bias in Graduate Admissions: Data From Berkeley", Science. 187 (4175): 398–404, 1975

~~Vicious Artificial~~ Enlightened Augmented Intelligence Cycle

ML combined with human insight

- Representative
- Reliable
- Plausibly causal

Well-designed data



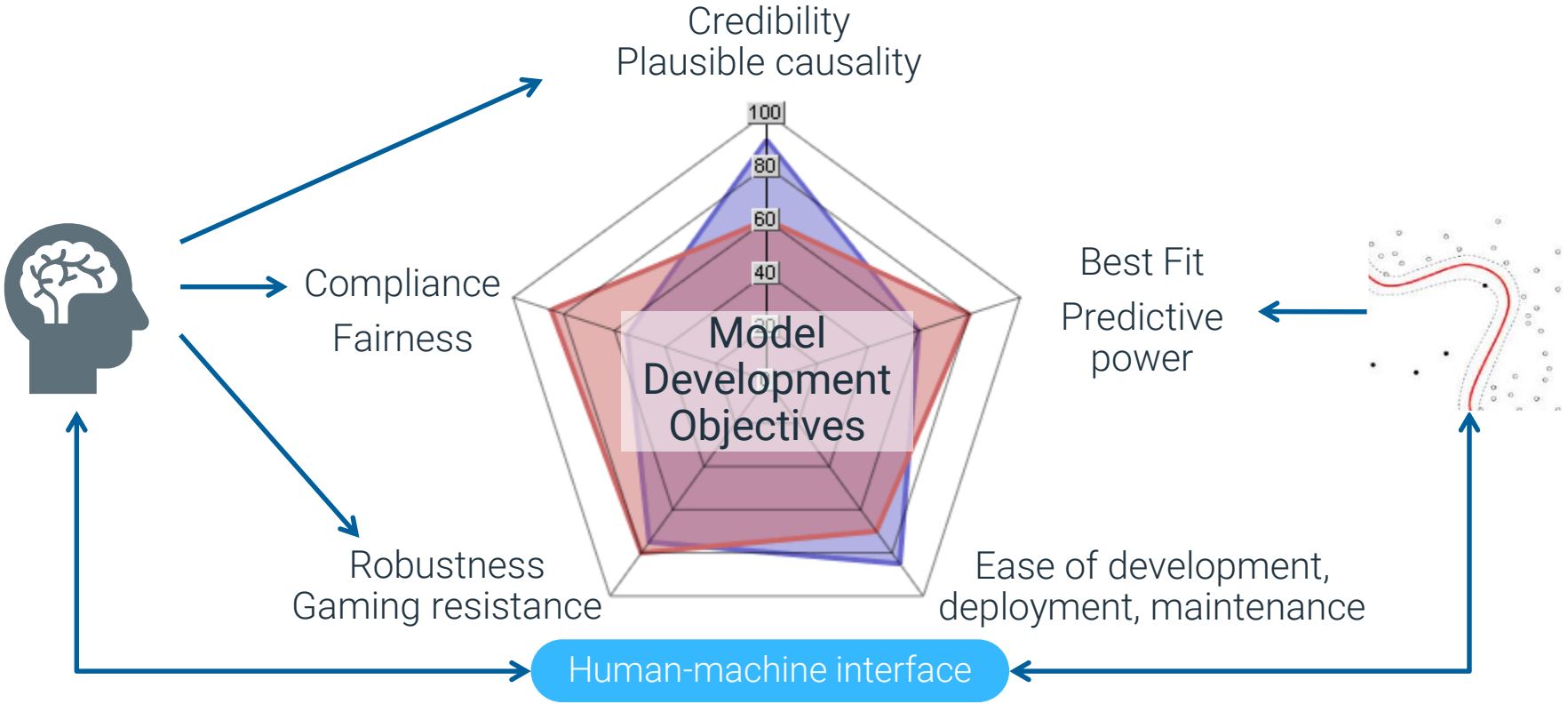
More palatable models

- Compliant
- Credible
- Plausibly causal

- Compliant
- Actionable
- Fair

Less biased, reasoned decisions, feedback to people

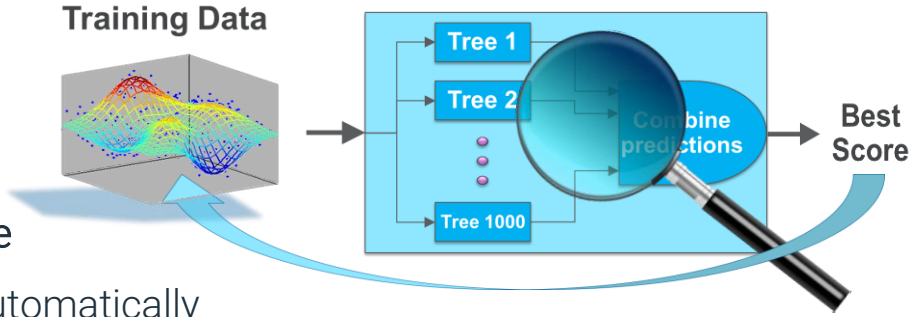
Bias Buster 2: Combine Machine Learning With Human Insight



Human-Machine Interface for Highly Effective Scorecard Development

https://www.thinkmind.org/index.php?view=article&articleid=data_analytics_2018_1_30_60077 (as of 10/17/2019)

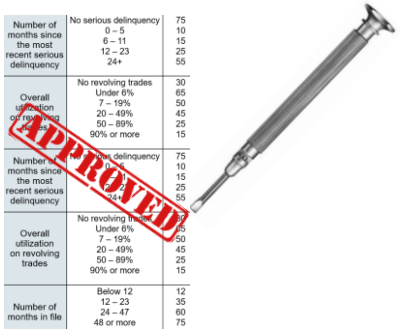
1. Develop "best fit" black box AI/ML model
2. Seek to understand black box
3. Augment data with **Best Score**
4. Approximate **Best Score** by automatically grown, explainable scorecard system



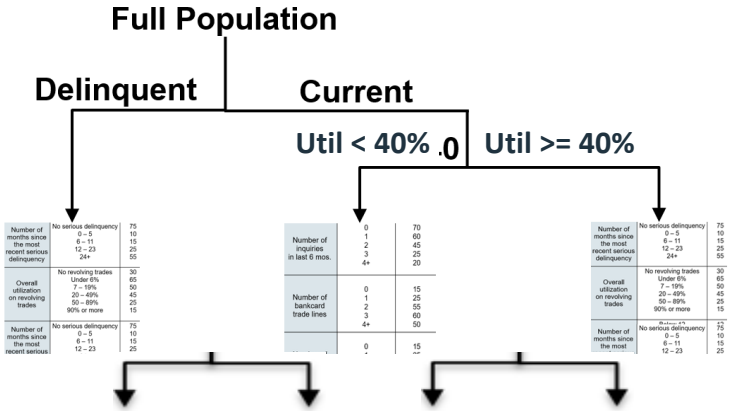
Diagnostics:

- Predictive performance benchmark
- Variable importance
- Nonlinear patterns, interactions

5. Hone final scorecards to warrant compliance



6. Deploy multi-scorecard system



Human-Machine Interface Dominates Other Approaches

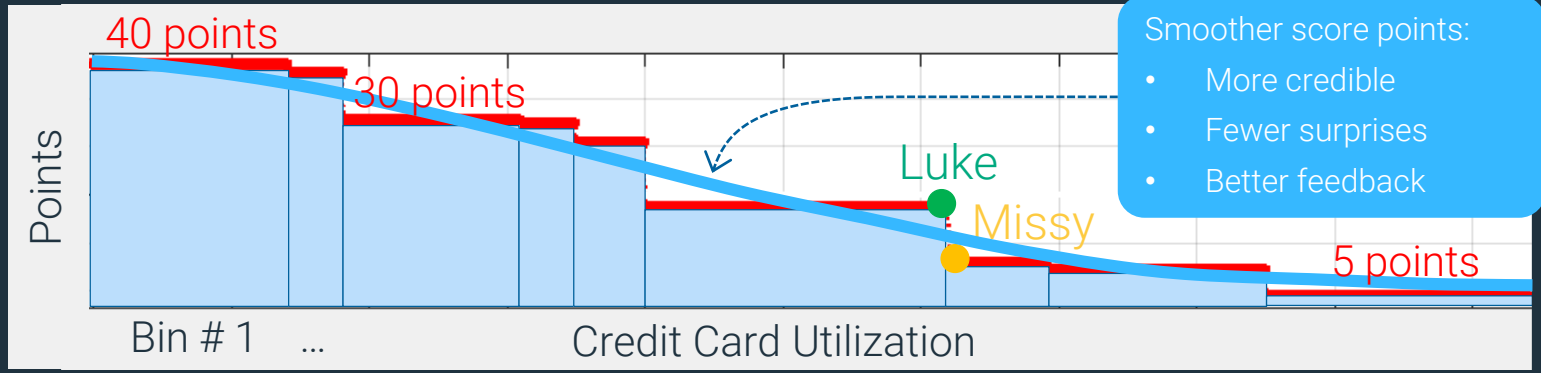
Model / Technology*	Development Effort (Resource Hours)	Explainability / Credibility	Predictive power improvement** over FICO® Score 9
FICO® Score 9	Laborious	High	N/A
Gradient Boosting	Modest	Limited	+1.7%
Multilayer Neural Net	Modest	Limited	+0.5%
Human-Machine Interface	Modest	High	+0.3%

*Same data used to train and test all models

**Performance metric: KS (Kolmogorov-Smirnov statistic)

Further Improving Explainability With “Liquid” Scorecards

Binning and Scoring of a Continuous Behavior Variable



	Credit Card Utilization	Score points
Luke	60%	18 Points
Missy	61%	8 Points

“Because the algorithm said there’s a bin break here” ?

Prudently Applied AI/ML for Credit Scoring Brings Benefits at Scale

Subjective decisions / Rampant discrimination

Credit scoring eliminates judgment

Regulation protects vulnerable groups

FICO® Score for diverse populations

Greater financial inclusion with alternative data (UltraFICO™ Score, FICO® Score XD)

We keep innovating to reduce bias and boost fairness

FICO® SCORE

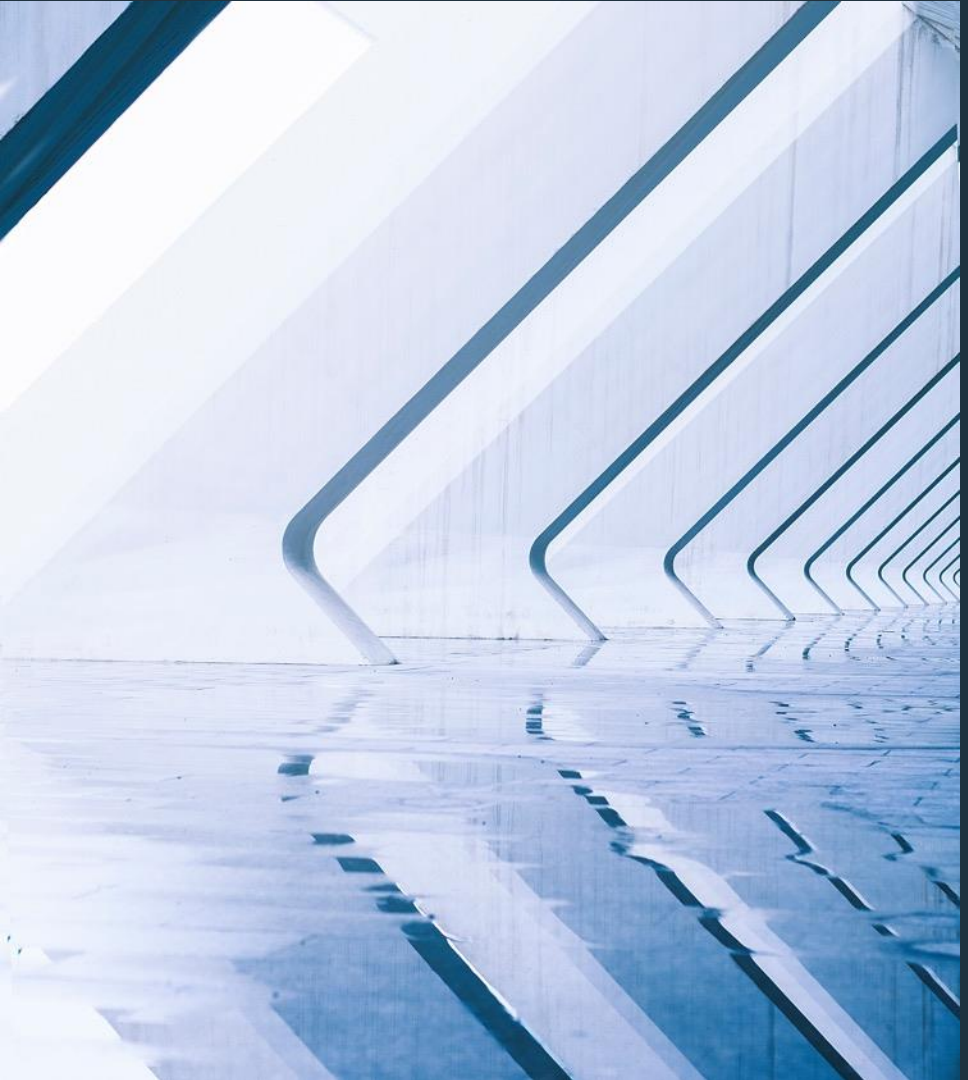
“Biased AI” is not an inherent flaw.

FICO® SCORE

Human and artificial intelligence can combine to fight biases at a massive scale.

If it sounds easy, it isn't.

But it's possible, and it's necessary.



Thank You!

Dr. Gerald Fahner